An investigational analysis of rs1061170 and its implication in Age-Related Macular Degeneration (AMD)

Authors: Lauren Schefter, Steven Dea, Jim Wright

INTRODUCTION

In the new era of genomics and personalized medicine, discoveries of new techniques and technologies are occurring at a rapid pace. Discoveries such as RNAi and CRISPR highlight the promise of potential therapeutics and new findings in our knowledge of disease. The underlying genetic factors are an important component in disease origin and development. Examples such as cystic fibrosis, Huntington's disease, beta-thalassemia, and sickle cell anemia illustrate simpler etiologies driven by mutations of only one gene, and even sometimes only a single base mutation. These single base changes are not always pathogenic, and are called SNPs, or single nucleotide polymorphisms. By contrast, it is becoming clearer that quite a number of pathologies are driven by multifactorial causes, meaning multiple genes and multiple mutations. In these cases, there is more than just the gene's function that may be impacted by mutation. Systems biology has revealed that these mutations' impact is found to be as significant for the disruption in overall homeostatic and development processes as it is in the one process or pathway of a single gene.[1] These larger integrations of gene expression are referred to as gene regulatory networks (GRNs). The GRNs that are often crucial to understanding of multi-factorial causes can be quite complex, with the example of several debilitating neurodegenerative pathologies such as Alzheimer's disease and Parkinson's disease. With this additional complexity, the promise of personalized medicine for multifactorial diseases remains strong but is perhaps a slightly more distant destination.

A good example of this multifactorial category of disease is age-related macular degeneration, or simply AMD (or even ARMD). The connection between this disease and genetics was strongly emphasized in a landmark 2005 GWAS study showing a positive genetic link in the risk and development of AMD.[2] The initial excitement led to more careful attempts at understanding as it became clear that AMD was clearly a multifactorial disorder. One of the initial genetic links was a SNP identified as rs1061170.[2] This variant can lead to a nonsynonymous mutation that results in the substitution of Histidine for Tyrosine at residue position 402 within the polypeptide sequence of the protein complement factor H (CFH). Thus, the mutant SNP allele is often referred to as Y402H. Considering rs1061170 has previously been implicated in AMD development, we attempted to determine what influence and impact this SNP has as an exonic missense mutation positioned in the protein coding sequence for CFH and its significance in the risk of developing AMD. While a single SNP may not be the primary determinant of pathology for multifactorial disease, it can often be a powerful method for elucidating the disease mechanism.

CFH Protein Structure and Function

In order to better understand and explore the Y402H mutation, a more detailed understanding of the CFH protein is needed. The *CFH* gene is responsible for generating the complement factor H protein. CFH is highly important in the body's regulation of the complement system, which is a key part of the innate immune system. The complement system involves several proteins, including CFH, that help coordinate innate immune responses. These proteins work together to trigger inflammation, remove cellular debris, and destroy bacteria/viruses.[3] CFH's main role in the complement system is to regulate the other proteins so that healthy cells are not destroyed or harmed by accident.

The way the complement pathway and more specifically, the alternative complement pathway works is that once foreign invaders are present, the body secretes multiple plasma and membrane-associated proteins in an attempt to protect the body. These molecules then bind to invading microorganisms and form protease complexes called C3-convertases.[4] These convertases then cleave the α-chain of C3, which then creates C3b. This cleavage exposes internal thiolester which allows C3b to bind to biological surfaces, thus exposing hydroxyl and amino groups.[4] After C3b has bound to these biological surfaces, it allows for phagocytosis by macrophages and destruction of the invading body. CFH comes into play when C3b levels need to be maintained at a lower level in the body when there is no microbial threat. CFH regulates the complement pathway in fluid-phase and cellular surfaces in three ways. The first way is that it can bind to C3b and cleave it, thus rendering it inactive. It can also accelerate the decay of the C3-convertase pathway as well as work as a cofactor with factor I to inactivate C3b2. CFH can also interact with CRP (C-reactive protein) and can inhibit the CRP pathway (damaged tissue induced pathway).

The *CFH* gene is located on chromosome 1q31.3 within the RCA gene cluster.[5] Figure 1 captures a view of this locus and surrounding genes.

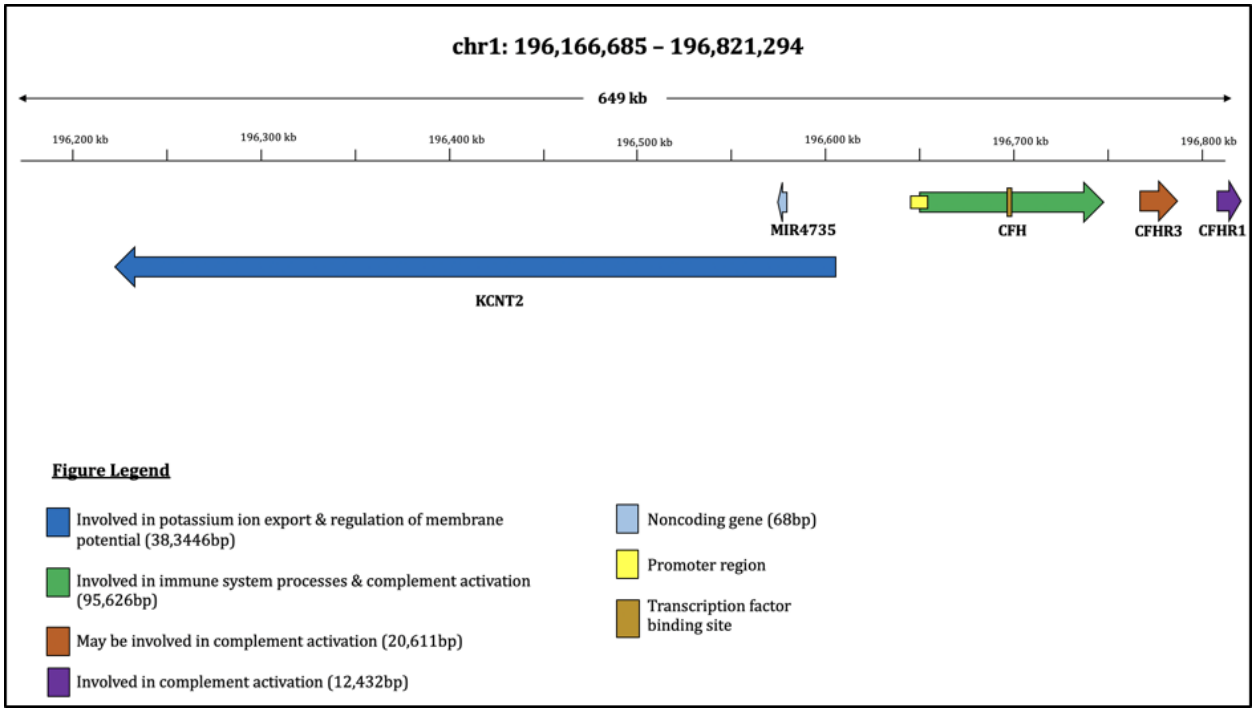


Figure 1: Arrows represent genes found on the Homo sapiens chromosome 1: 196,166,685 - 196,821,294. Arrow direction is representative of strand direction; the right pointing arrow indicates the gene is on the forward strand while the left pointing arrow indicates the gene is on the reverse strand. In addition to the genes themselves, promoter and transcription factor binding site (TFBS) are elucidated as well.

The canonical transcript for the *CFH* gene is made up of 22 exons, but there are five alternatively spliced isoforms with fewer exons but of similar protein structure (Figure 2). One pertinent example for the ocular environment (which will be examined later in more detail) produces a protein product called FHL-1. The protein CFH is produced in the liver constitutively and can be found in human plasma at concentrations of 220–650**µ**g/ml.[8] Other than the liver, various types of cells also produce CFH including peripheral blood lymphocytes, myoblasts, rhabdomyosarcoma cells, umbilical vein endothelial cells, glomerular mesangial cells, neurons and glial cells.[4] The ability of the body to produce CFH not only in the liver, but also in these other cells allows for better regulation of the complement system when performing correctly.

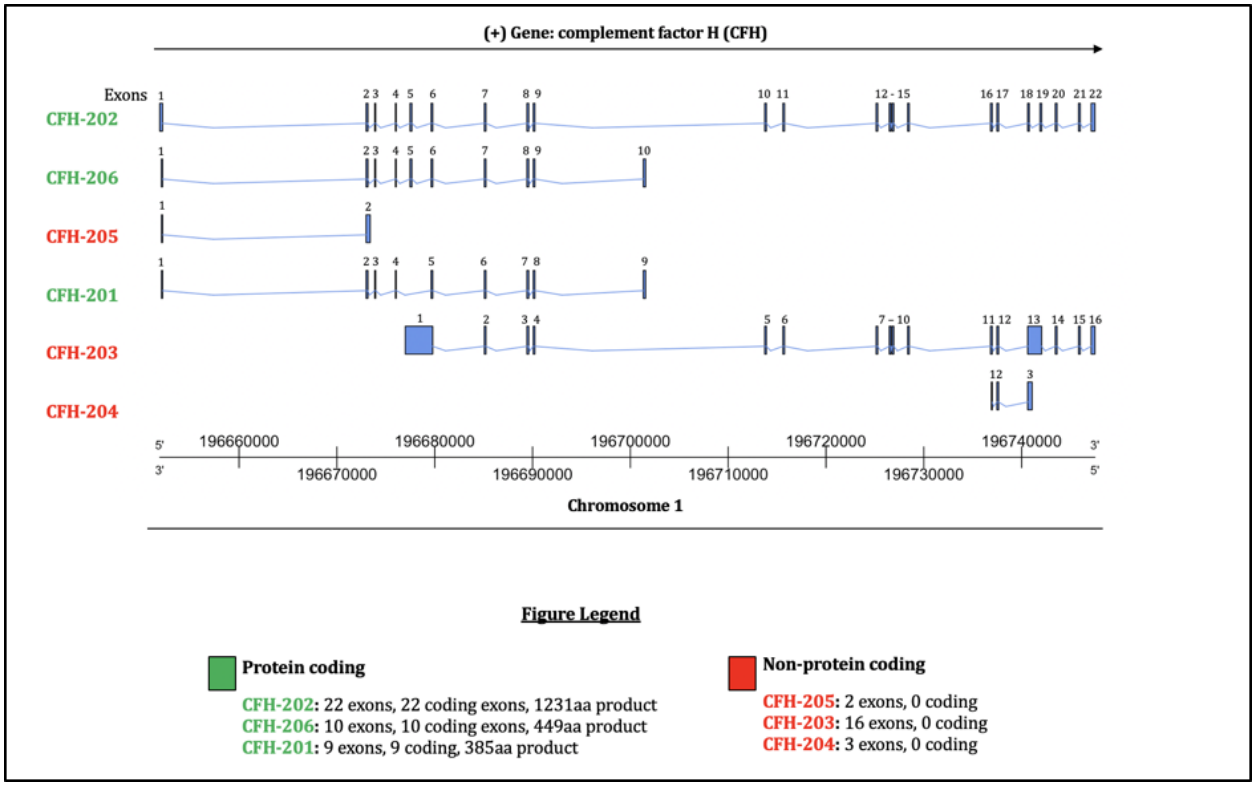


Figure 2: Represents the complement factor H (*CFH*) gene found on forward strand of chromosome 1 from position 196,651,878 - 196,747,504. This gene has six transcript variants, three are coding (CFH-202, CFH-206, CFH-201) and three and non-coding (CFH-205, CFH-203, CFH-204).

Genomic Foundation for SNP rs1061170 and Corresponding Y402H Variant of the CFH Protein

The *CFH* gene is found on chromosome 1 and occupies over 95-kbp, but it is actually a part of a larger CFH gene family locus comprising approximately 360-kbp. This *CFH* gene family consists firstly of *CFH*, this being the ancestral gene, and then five other *CFH*-related genes known as *CFHR1* through *CFHR5*. The *CFH* gene family have significant structural similarity (Figure 3)[6,7] given their evolutionary origins due to segmental duplications (SDs) that have occurred during primate evolution.

In a recent study (Cantsilieris et al. 2018) these SDs were analyzed and compared across six primate species (including humans). Eight distinct SDs were identified ranging in size from ~2.5-kbp to 30-kbp.[7] The authors estimate the *CFH* family locus has expanded ~threefold over 40 million years driven almost exclusively by segmental duplication (SD). Essentially the five *CFH*-related genes now present were derived (at different evolutionary timepoints) from incomplete duplication events that still contained protein-coding exons with respect to the ancestral *CFH* gene.

As stated previously, the rs1061170 variant within the coding sequence of the *CFH* gene (exon 9) results in a residue substitution of Histidine^{402} for Tyrosine^{402}. The CFH protein consists of twenty short consensus repeat (SCR) domains, and the Tyr^{402} → His^{402} variant (Y402H) occurs in SCR7. It has been shown that SCR6 to SCR8 domains in CFH form a heparan-sulfate binding site responsible for CFH binding to cell surface GAGs (glycosaminoglycans) as part of identifying self (host) cells in the macula. This CFH binding protects said cell by identifying self from non-self and prevents an autoimmune response.[8] Additionally, CFH domains SCR19 to SCR20 are also involved in GAG binding as a means of identifying "self".[8] The Y402H variant has been shown to result in decreased heparan-sulfate binding ability for the SCR6-8 domains.[9] Nonetheless, the principally recognized function of CFH protein is as a negative regulator of complement activation, serving to downregulate the alternative pathway of the complement system. Figure 3 provides a view of the SCR domains within the CFH protein and *CFH*-related gene products.

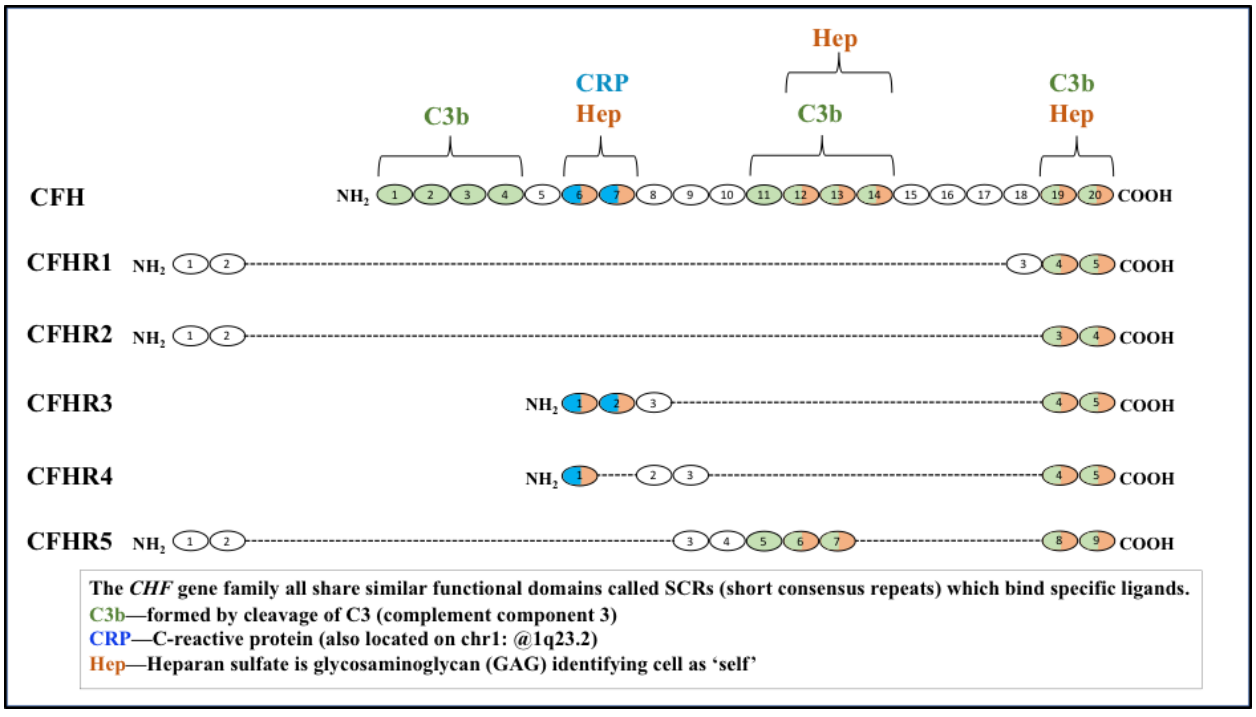


Figure 3: *CFH* gene family structural similarity: CFH protein comprises twenty SCR domains and the *CFHR* genes all possess some portion of these.[6]

Cantsilieris et al. designate one particular segmental duplication as SD-A, which overlaps with exon 8 and exon 9 of the *CFH* gene.[7] In fact, they determined that *CFH* exons 8 and 9 are part of a duplication cassette that has been recurrently reused in at least five separate SD events. It is interesting to note that rs1061170 is found within exon 9, and affects the SCR7 domain (Figure 4). The SD blocks cause the *CFH* family locus to be subject to unequal crossing over and gene conversion. The study's sequence analysis also shows presence of Y402H among the lesser apes, which implies that the Y402H variant has been present very early in the ape lineage (~20 Mya).

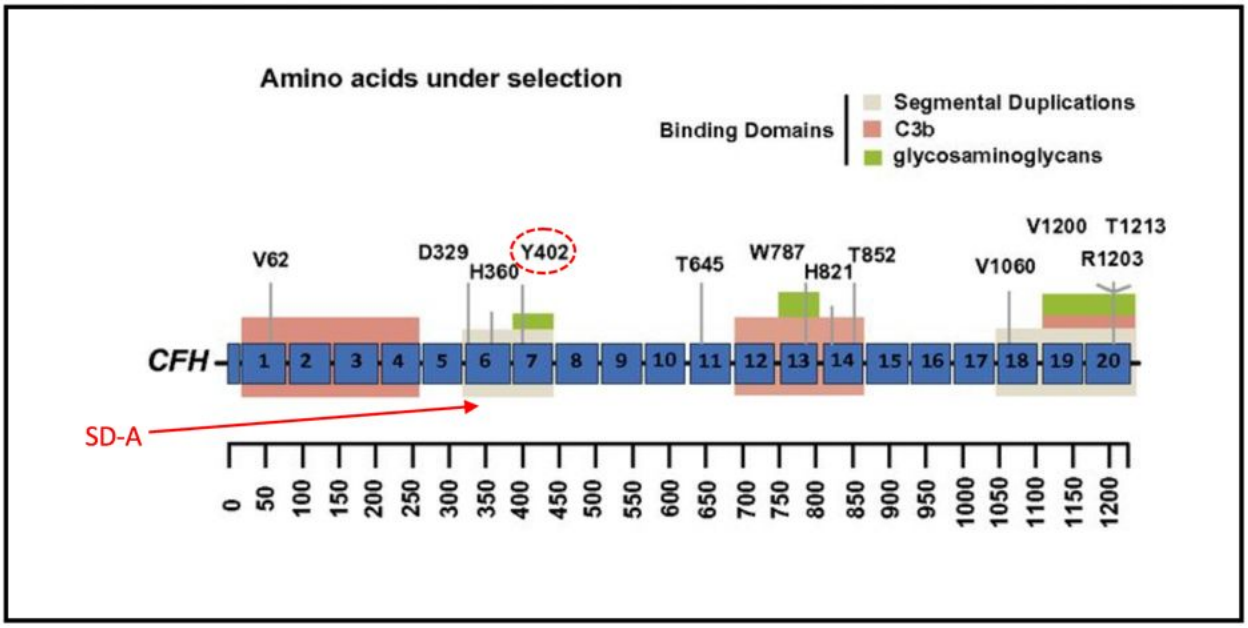


Figure 4: Diagram illustrating position of segmental duplications (grey) within *CFH* SCR domains.[7]

The missense mutation that occurs with variant Y402H involves a change to base 1277 within exon 9 from a T to a C (so-called T1277C).[10] Three genotypes exist for this variant: homozygous CC or heterozygous CT (both mutants), or the non-mutant TT genotype. The variant Y402H exists in only four of the six CFH isoforms where the exon it affects is present (Figure 5). Our variation of Y402H is found in exon 9 of the "normal" CFH transcript and can be seen at its location in the three other isoforms of CFH in Figure 6.

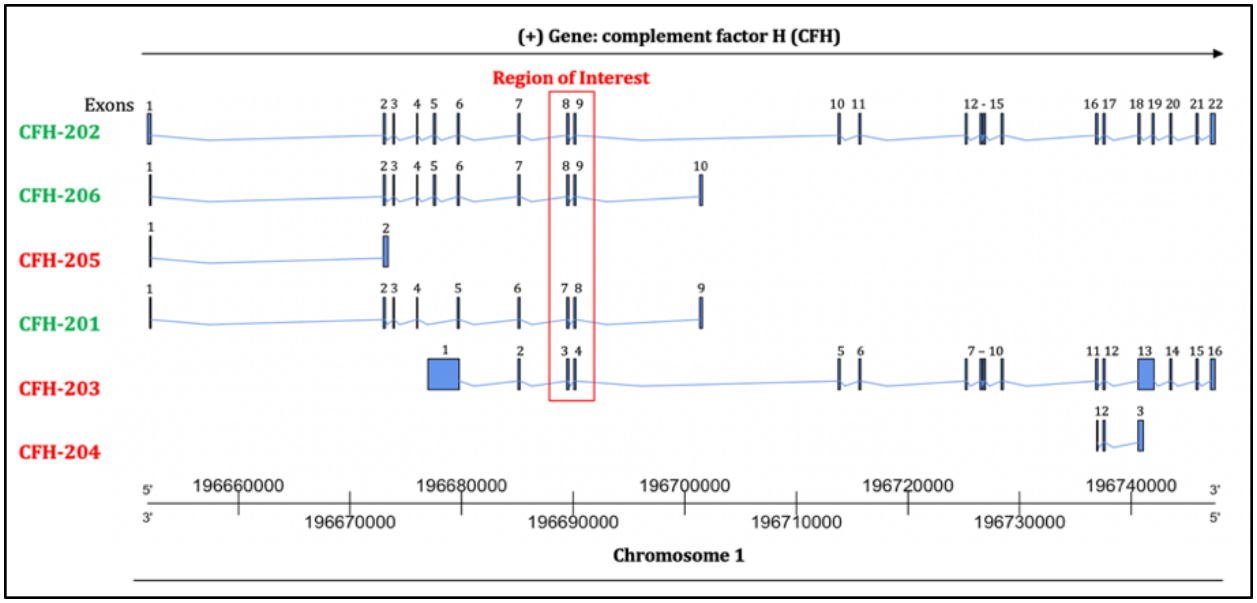


Figure 5: Diagram illustrating position of SNP rs1061170 within the six CFH isoforms.

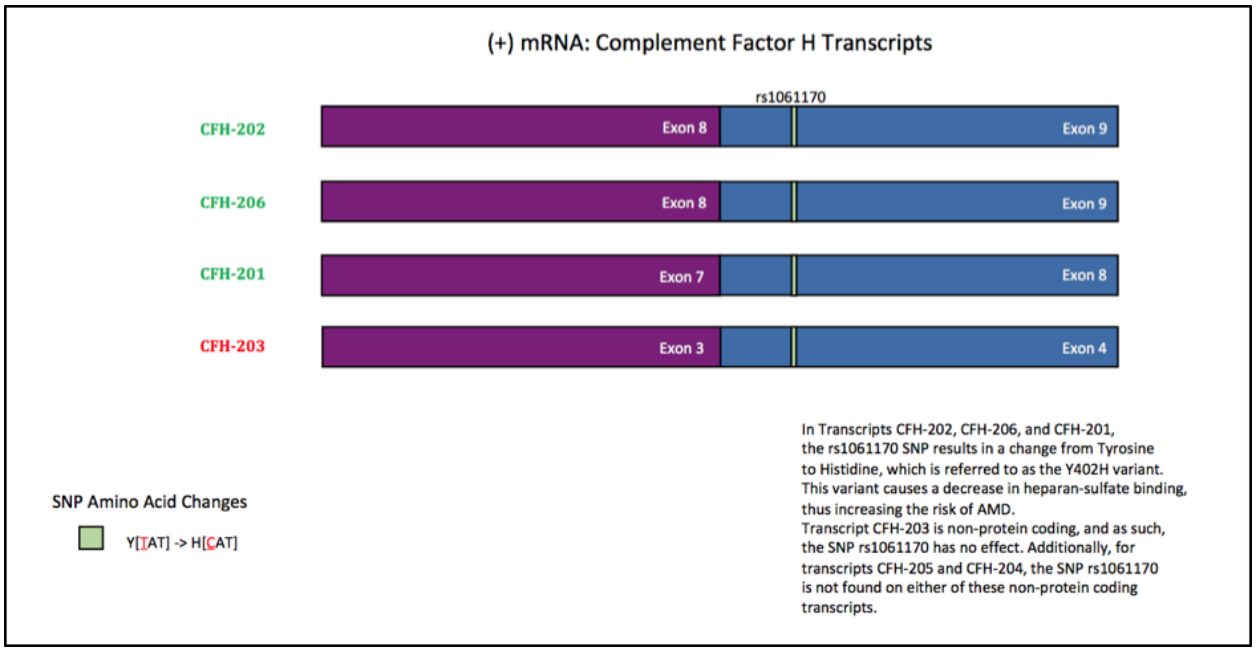


Figure 6: This diagram shows rs1061170 position in CFH transcripts.

The 2005 GWAS study based on genetic linkage evidence in AMD patients was directed towards finding genetic factors and identified an intronic SNP (rs380390) having strong association with AMD.[2] In the same study the authors further identified the functional variant later known as Y402H (rs1061170). Since then over 24,000 SNPs have been identified within the *CFH* gene.[11] However, only 13 are identified as pathogenic. Some of the more important SNPs that display an association with AMD risk are rs2274700, rs1410996, and rs7535263, in addition to our own SNP: rs1061170 (Figure 7).

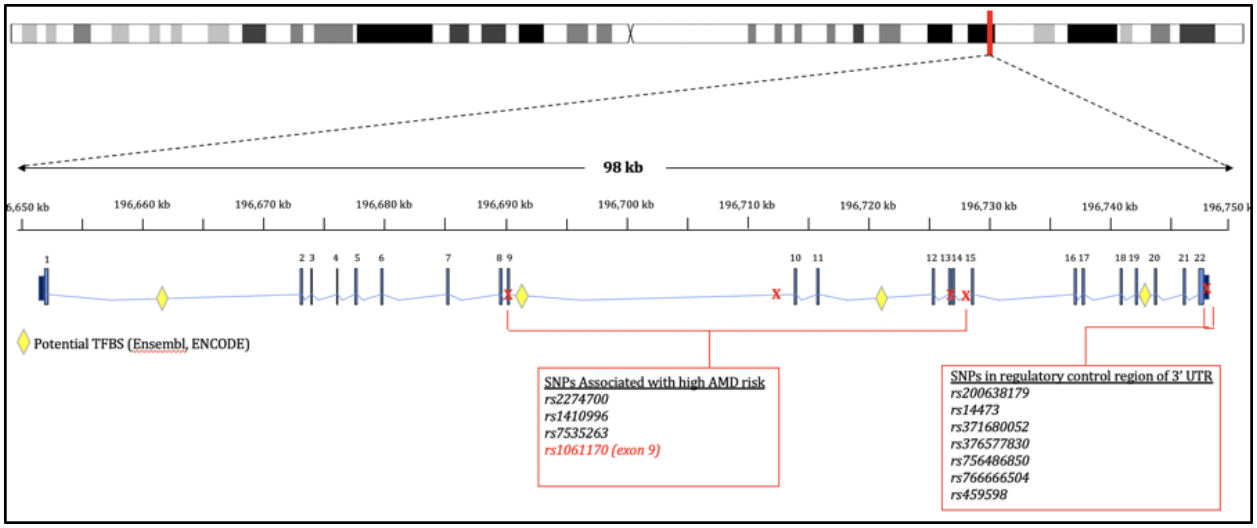


Figure 7 Diagram illustrating position of SNP rs1061170 along with other SNPs at *CFH* locus.

Pathology for Age-Related Macular Degeneration (AMD)

The pathology for AMD is a slowly progressive degeneration of the macula of the retina, and is generally divided into the early, intermediate, and late stages. Detection of the condition in early and intermediate stages involves imaging for the presence of lipid-protein deposits known as “drusen”. Drusen particles have also been found to have high concentrations of complement system proteins, including CFH. There are two basic types of late stage AMD, one known as wet AMD, also known as neovascular AMD (nAMD), and the other type is known as dry AMD, which broadly refers to any late stage AMD that is not neovascular. The wet AMD disease type typically occurs in approximately 10% of all cases. There are currently no successful therapeutics for dry AMD.

The macula is a region with a higher density of photoreceptors cells than anywhere else in the retina. This region is therefore responsible for high-acuity vision and also contains the concentrated region of color photoreceptor cells, called cone cells, known as the fovea. Degeneration and damage of the macular region results in degrees of vision impairment that can

progress from "fuzzy" vision to the point of being effectively blind. Progression of vision impairment tends to be slower in patients with dry AMD versus wet AMD. The ocular region of AMD pathogenesis is the space near the outer retina containing the retinal pigmented epithelium (RPE), Bruch's membrane, and the choroid vascular layer nourishing retinal cells and tissue (Figure 8). The drusen deposits are usually detected within Bruch's membrane (BM) or between the RPE layer and BM.

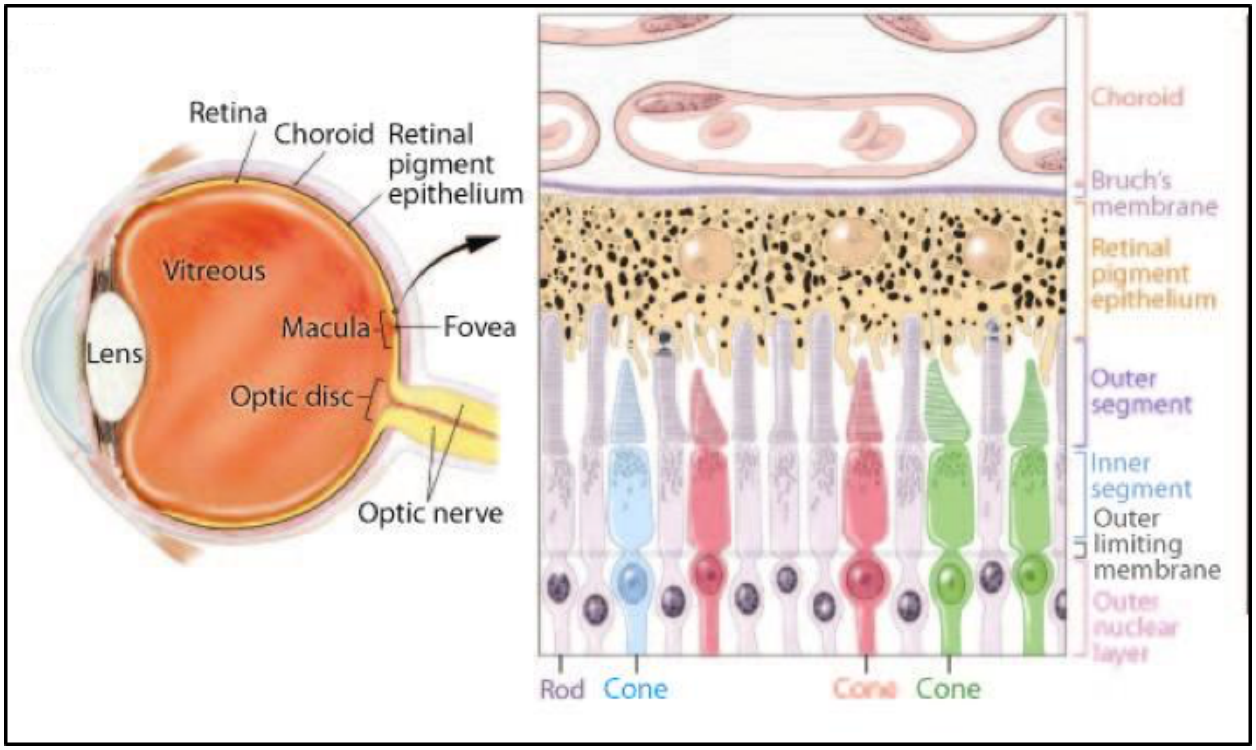


Figure 8: This diagram reveals ocular anatomy including the macula, RPE, and choroid layers. The inset provides a more detailed view showing photoreceptors, RPE, and Bruch's membrane adjacent to the choroid layer with vascular vessels in cross-section.[12]

CFH Link to AMD Pathobiology

Recent studies have shown that dysfunction associated with CFH is a major factor in development of AMD. It has been shown that CFH and lipoproteins compete for binding in the sub-RPE. With a decrease in CFH levels observed, the lipoproteins then form deposits in the sub-RPE (as there is no longer as much competition for binding), and this causes the complement system to activate and do damage to the RPE, potentially leading to vision loss.[13]

In this case, CFH regulates the amount of lipoproteins that are able to form deposits in the RPE, rather than directly influencing the activity of the complement system.

A more recent line of investigation in AMD pathobiology involves a protein called FHL-1. An alternative *CFH* gene product, the mRNA splice isoform known as factor H-like protein 1 (FHL-1), is also subject to the missense mutation of Y402H. It's possible that the

FHL-1 isoform is a more likely culprit as a primary factor in AMD development due to its more active role in the ocular environment (see details below). FHL-1 is a smaller protein (~49-kDa versus the 155-kDa of full-length CFH protein). Thus, FHL-1 is comprised of only the first seven SCRs of CFH (not the full twenty). Given that CFH has two separate heparan-sulfate binding domains, whereas FHL-1 has only the one via SCR7, it seems reasonable that FHL-1 variant (Y402H) function might be more impacted than CFH by Y402H. Several studies[8] have shown a strong association of increased complement activation with risk and development of AMD, and the Y402H variant has been considered for years as a major risk factor for AMD.

Recent research has shown that FHL-1 is able to diffuse across Bruch's membrane whereas CFH is not, making FHL-1 regulatory function possibly even more important in the sub-RPE space where AMD pathobiology is thought to start.[14] This study examined the diffusion of CFH and FHL-1 across Bruch's membrane[14] and found that while FHL-1, with only seven SCR domains, passively diffuses easily across the membrane, CFH protein (with 20 SCRs) was not found on the RPE side of the membrane (i.e., did not diffuse across). Bruch's membrane is essentially the barrier between the blood supply in the ocular choroid region and the RPE layer that replenishes the retinal photoreceptors. The authors in this study determined that the main protein regulator of complement activity in the RPE-Bruch's membrane region is FHL-1, not CFH.[14] SNP rs1061170 is found in exon 9, and it is now clear that Y402H directly affects the 7th SCR domain (Figure 9).[14] FHL-1's last C-terminal domain is SCR7, directly followed by a 4-residue C terminus. The extracellular matrix (ECM) character of Bruch's membrane means that a regulator protein like FHL-1 is not in soluble form, but rather is bound to heparan sulfate sites. The Y402H variant directly impacts FHL-1's ability to successfully bind heparan sulfate in ECM. Unlike CFH protein, which has two separate heparan sulfate binding domains 17(SCR7 and SCR19-20), FHL-1 has only one and so is more likely to be significantly affected by SNP rs1061170. Thus, there is potential failure of regulatory control of complement activation in the Bruch's membrane-RPE region that could potentially lead to development of AMD.

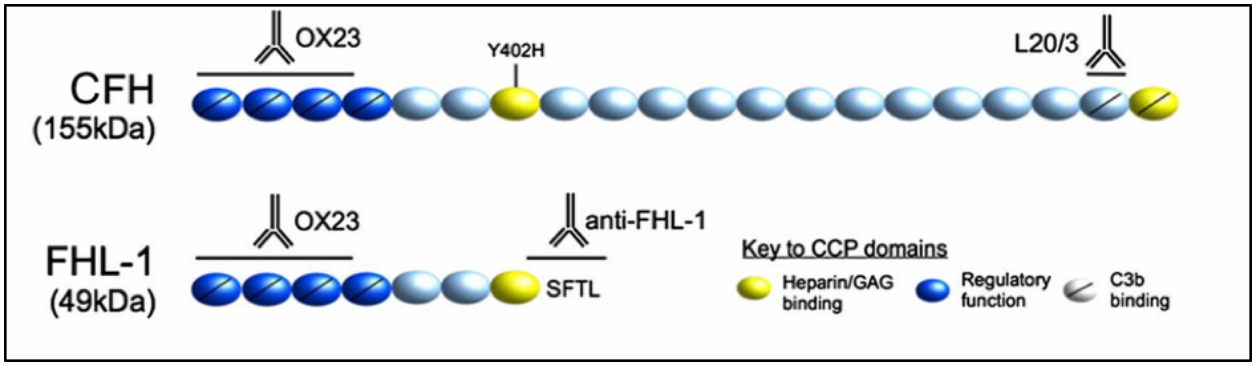


Figure 9: This diagram highlights the different structures for CFH vs. FHL-1 and the difference in impact of the presence of Y402H variant.[14]

The exact nature of how Y402H affects heparan sulfate binding is not well understood. Protein crystalline studies of $SCR7^{H402}$ and $SCR7^{Y402}$ show very similar conserved structures. However, in one study utilizing computational simulations of molecular dynamics, the authors test their hypothesis that differences in side-chain flexibility between the two isoforms affect ligand binding.[15] Figure 10 illustrates their hypothesis (orange is the H402 variant).

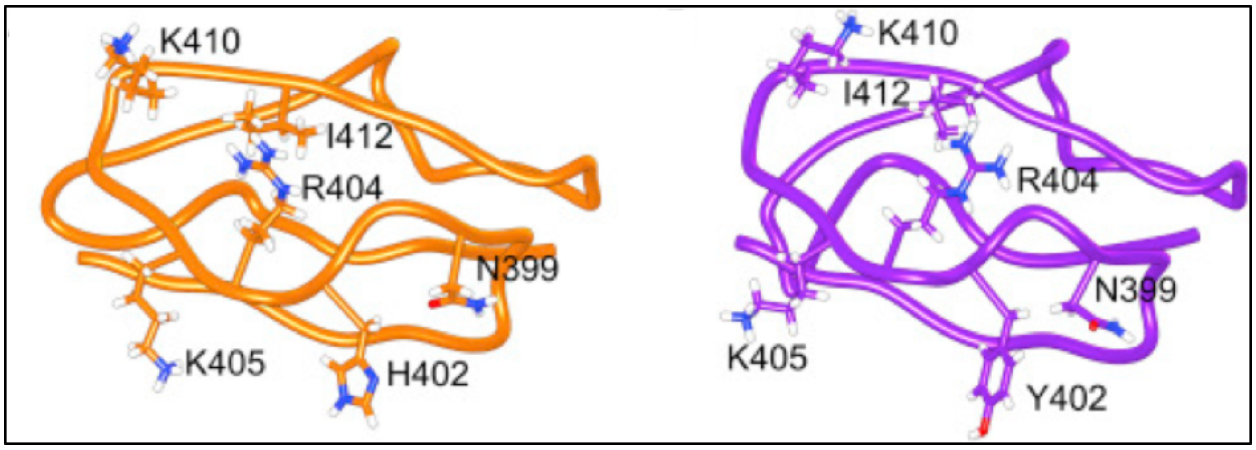


Figure 10: Comparison of the different molecular structures for SCR7 domain with H402 (mutant in orange) versus SCR7 with Y402 (purple).[15]

METHODS

The following methods and procedures were performed to generate and retrieve data using a variety of resources for our SNP investigation of the *CFH* gene:

BioMart & Ensembl Analysis for the Investigation of TFBS, ncRNAs & CNVs

We performed a biomaRt query (getBM) using the dataset Human Regulatory Evidence (GRCh38.p12) with filters— Chromosome: 1, Start(bp): 196690000, End(bp): 196720000, Feature Type, Feature Type Class: Transcription Factor.

Another biomaRt query used dataset Human genes (GRCh38.p12) with filters — Chromosome:1, Start(bp): 196600000, End(bp): 196850000; and attributes — RefSeq ncRNA ID, RefSeq ncRNA predicted ID, Transcript start (bp), Transcript end (bp), Transcript length (including UTRs and CDS), and Transcript stable ID.

As a starting point, Ensembl was used to look for variants within the *CFH* gene. Within the Genetic Variation category is the Variant Table which displays short sequence variations within the *CFH* gene. If the same variant falls in several transcripts within the same gene, a new row is created for each transcript.[16]

BioMart was used next to confirm the results previously found in Ensembl and gain additional insight concerning the SNPs and CNVs found in the *CFH* gene region. After selecting the Ensembl Variation database and Human Short Variants dataset, the same information was found using chromosome 1, 196,651,878 as the start position, 196,747,504 as the end position and clinical significance as the filter (specifically "pathogenic", "pathogenic, risk factor" and "benign,pathogenic,risk factor" were selected).

Galaxy & IGV Visual Analysis to Compare SNPs & CNVs in Region of Interest

Galaxy was used to compare the information on SNPs and CNVs previously reported by getting information from the UCSC Table Browser. To get the equivalent information, the hg38 genome assembly, all SNPS(151) track, and CFH-202 transcript id position (chr1:196651878-196747504) were selected.

The Integrative Genomics Viewer (IGV) allowed for a more zoomed in view of both the *CFH* gene and the region directly surrounding the SNP of interest. After selecting the hg38 genome, loading the all SNPs track and common SNPS track from the server, and zooming in to chromosome 1, exon 9, from base pair 196,690,055-196,690,243 the results in figure 12 were found.

We chose to take a closer look at histone modification H3K27Ac in regards to the *CFH* gene as it is known to be a key histone marker for transcriptional activation, or activation of enhancers.

NGS Sample Selection & Analysis

Using results from these queries as a guide, we then selected next generation sequencing (NGS) transcriptome (RNA-seq) sample data and ran a bioinformatics pipeline to determine presence of our SNP variant. The control sample (SRR5488335) was taken from a transcriptomic sample set in a 2018 gene expression study on human eye development (PRJNA384924, Cambridge U). Both AMD samples (SRR5601894, SRR5601895) are taken from a 2018 study on chromatin accessibility in AMD (PRJNA388006, Johns Hopkins U).

Our main goal in selecting two data sets was to find both control and AMD-positive samples to run through our next generation sequencing pipeline. After attempting to find samples from the same experiment in order to minimize biasing factors (e.g., sample preparation) in our analysis, as well as correlating traits like age and gender, the variation among protocols hindered finding acceptable samples. Eventually we were able to choose three RNA-seq samples (described in Table 1) that clearly denoted whether or not they were a control or an AMD-positive sample and proceeded to run our NGS analysis. It should be noted, however, that the control sample was run on a different Illumina platform than the AMD samples as it was difficult to find data to fulfill every desired requirement. We chose samples that came from retinal tissue, RNA-seq analysis, and had detailed information about the donors. Fortunately, the differences between the MiSeq and NextSeq-500 Illumina platforms are fairly negligible as the

main difference lies in the number of samples or the size of the samples possible to be sequenced. Thus, we felt this was a detail that would least impact our analyses.

Both the control (SRR5488335) and AMD samples (SRR5488335 & SRR5601894) were obtained from retinal tissue. Both AMD samples were obtained from one individual post-mortem. The right eye (AMD sample 1) was characterized as early AMD, while the left eye (AMD sample 2) was identified as late stage AMD. Following collection, RNA was extracted using QIAGEN RNeasy kits per the manufacturer's protocol. RNA integrity was accessed for all samples before proceeding to library preparation. AMD libraries were assessed by DNA-based fluorometric assay and automated capillary electrophoresis, however, it is unclear as to if this was also performed on the control sample library.

	Control Sample	**AMD Sample 1**	**AMD Sample 2**
SRR	SRR5488335	SRR5601894	SRR5601895
Biosample	SAMN06854061	SAMN07166539	SAMN07166538
Sex	male	male	male
Age	83	87	87
Disease	none	Early AMD	Late AMD
Tissue	retina	retina	retina
Platform/Model	Illumina NextSeq 500	Illumina MiSeq	Illumina MiSeq
Library Layout	Paired	Paired	Paired
Library Strategy	RNA-seq	RNA-seq	RNA-seq
Read Count	62,386,081	62,644,363	83,916,918
Base Count	9,415,786,999	12,528,872,600	~16,800,000,000

Table 1 Comparison of selected NGS sample metadata

RESULTS

Transcription Factors

When we ran our first query with a range from 196690000 to 196720000, it returned over 300 results. A reduced query range (Start(bp): 196690000, End(bp): 196700000) of 10kb was run – and this yielded a more convenient list of 10 TFs (Table 2).

Chromosome	Start (bp)	End (bp)	Length (bp)	Feature type	Feature type class
1	196690749	196691068	319	Max	Transcription Factor
1	196690764	196691026	262	MYC	Transcription Factor
1	196692182	196692813	631	ZZZ3	Transcription Factor
1	196692257	196692929	672	ZNF639	Transcription Factor
1	196696305	196696717	412	NFIC	Transcription Factor
1	196697508	196697836	328	NFIC	Transcription Factor
1	196698993	196699348	355	CEBPB	Transcription Factor
1	196699038	196699345	307	CEBPB	Transcription Factor
1	196699039	196699315	276	CEBPB	Transcription Factor
1	196699053	196699324	271	CEBPB	Transcription Factor

Table 2 Transcription factor list as a result of a BioMart query limited to the range of base pairs 196,690,000 - 196,700,000.

When we viewed the TF coordinates highlighted above in Ensembl's genome browser with Regulation features displayed, there is a transcription factor binding site called NFIC that has genomic coordinates of roughly the same length and is located roughly 36 k-bp away (Figure 11). NFIC, or nuclear factor 1 C-type, is a common dimeric DNA-binding factor. Due to proximity, this would be an interesting area for future investigation as there currently is no additional data pertaining to an association with the SNP rs1061170.

Source	Start (bp)	End (bp)	Length (bp)	Feature type *or* Ensembl ID
biomaRt	196696305	196696717	412	NFIC
Ensembl	196659681	196660120	439	ENSR00000955312

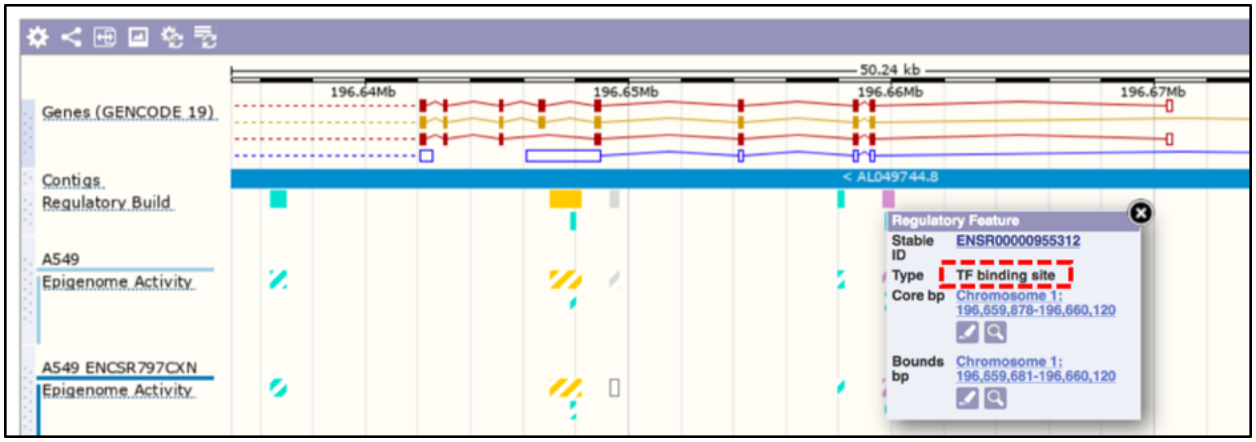


Figure 11: Potential Transcription Factor binding site (TFBS) identified in Ensembl

Noncoding RNA

The predicted (XR_) ncRNA transcripts as a result of the BioMart query (limited to our region of interest) are either transcript variants for CFH (XR_001737134) or for other adjacent genes in the same region (CFHR3: XR_426757,XR_001736938, XR_002958987, XR_001736937, XR_241062; and ZBTB41: XR_002956436). The ZBTB41 variant is also the only annotated transcript variant with the accession NR_135153.1. The annotation references three published studies (PMIDs: 24925725, 23381943, and 16341674), and the refseq entry has been validated. The variant is designated noncoding RNA because the expected start codon renders the transcript a candidate for nonsense-mediated RNA decay (NMD) (Table 2).

RefSeq ncRNA_ID	RefSeq ncRNA predicted ID	Transcription start (bp)	Transcription end (bp)	Transcript length	Transcript stable_ID
		197138748	197139307	560	ENST00000442280
		196819757	196832189	1271	ENST00000320493
		196819781	196832189	1070	ENST00000367424
		196819822	196821202	399	ENST00000472961
		196819865	196826026	588	ENST00000468079
		196825224	196828167	723	ENST00000480960
	XR_001737134	196651878	196747504	4127	ENST00000367429
		196652045	196701566	1658	ENST00000630130
		196652056	196673407	550	ENST00000496761
		196652056	196701565	1454	ENST00000359637
		196676988	196747504	6985	ENST00000466229
		196736908	196741086	753	ENST00000470918
NR_039888		196582413	196582481	69	ENST00000580028
		197222222	197223255	1034	ENST00000417716
		197038741	197055823	1378	ENST00000649282
		197039191	197067267	2217	ENST00000367412
		197040145	197050845	596	ENST00000490002
		196850283	196884793	614	ENST00000649395
		196888014	196918713	2178	ENST00000367416
		196888052	196918633	2063	ENST00000608469
		196888082	196918633	1292	ENST00000251424
		196914750	196916630	478	ENST00000647770
		196888103	196918428	1066	ENST00000367418
		197437976	197447469	548	ENST00000422250
		197084128	197146694	10887	ENST00000367409
		197084132	197135144	3747	ENST00000367408
		197084128	197146694	6132	ENST00000294732
		197153680	197200542	8478	ENST00000367405
NR_135153	XR_002956436	197158442	197200521	3779	ENST00000467322
		197363817	197364078	262	ENST00000436696
		196977556	197009674	2810	ENST00000256785
		196943756	196959212	1325	ENST00000367421
		196943768	196959226	1072	ENST00000367415
		196943772	196959216	896	ENST00000649283
		196943772	196959226	1020	ENST00000476712
		196943776	196945675	917	ENST00000647617
		196943784	196949873	574	ENST00000485647
		196943817	196950939	424	ENST00000489703
		196943837	196959219	801	ENST00000496448
		196943858	196959215	599	ENST00000473386
		196943885	196959106	658	ENST00000649960
		196774795	196794073	1645	ENST00000367425
	XR_426757	196774813	196790022	1965	ENST00000471440
	XR_001736938	196774813	196790022	1965	ENST00000471440
	XR_002958987	196774813	196790022	1965	ENST00000471440
	XR_001736937	196774813	196790022	1965	ENST00000471440
	XR_241062	196774813	196793488	1242	ENST00000367427
		196774816	196793675	1043	ENST00000391985
		196789760	196793633	785	ENST00000461558
		196774816	196795406	1462	ENST00000617219

Table 2 Biomart query for noncoding RNA

SNPs and CNVs

When we looked at the Ensembl Genetic Variation category without refining the output table, 80,110 variants were found. Table 3 shows how many variants were found in each transcript of the *CFH* gene. Additionally, the SNP of interest (rs1061170) was found in four of the six transcripts: CFH-201, CFH-202, CFH-203 & CFH-206 (Table 3).

Name	**Transcript ID**	**bp**	**Protein**	**Biotype**	**Total Variants**
CFH-206	ENST00000630130.2	1658	449aa	Protein coding	13,624
CFH-202	ENST00000367429.8	4127	1231aa	Protein coding	26,340
CFH-201	ENST00000359637.2	1454	385aa	Protein coding	13,623
CFH-205	ENST00000496761.1	550	No protein	Processed transcript	5,534
CFH-203	ENST00000466229.5	6985	No protein	Retained intron	19,787
CFH-204	ENST00000470918.1	753	No protein	Retained intron	1,202

Table 3 CFH transcript variant data according to Ensembl

If the information is then limited to only include the variants in the transcript with a CCDS (CFH-202/ENST00000367429.8), and further filtered to only include those that are pathogenic, 18 variants remain. These pathogenic SNPs can be seen in Table 4 and the SNP of interest, rs1061170, is highlighted in yellow.

	Variant ID	vf	Location	Chr: bp	vf_allele	Alleles	Class	Source	Clin. Sig.	Conseq. Type	Transcript
2	rs387906550	506467169	1:196673968	1:196673968	G	T/G	SNP	dbSNP	pathogenic	splice region variant~intron variant	ENST00000367429.8
3	rs121913058	503756689	1:196676018	1:196676018	A	G/A/C/T	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
4	rs121913058	503756689	1:196676018	1:196676018	C	G/A/C/T	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
5	rs121913058	503756689	1:196676018	1:196676018	T	G/A/C/T	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
6	rs121913054	503756685	1:196677613	1:196677613	A	G/A/T	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
7	rs121913054	503756685	1:196677613	1:196677613	T	G/A/T	SNP	dbSNP	pathogenic	stop gained	ENST00000367429.8
8	rs796052138	513431071	1:196679674	1:196679674-196679676	-	AGA/-	deletion	dbSNP	pathogenic	inframe deletion	ENST00000367429.8
9	rs1061170	501974075	1:196690107	1:196690107	T	C/T	SNP	dbSNP	benign~pathogenic~risk factor	missense variant	ENST00000367429.8
10	rs121913061	503756692	1:196690125	1:196690125	T	C/T	SNP	dbSNP	pathogenic	stop gained	ENST00000367429.8
11	rs121913056	503756687	1:196690194	1:196690194	A	T/A/G	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
12	rs121913056	503756687	1:196690194	1:196690194	G	T/A/G	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
13	rs121913052	503756683	1:196715679	1:196715679	C	T/C	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
14	rs1131690796	527033563	1:196728506	1:196728506	-	A/-	deletion	dbSNP	pathogenic	frameshift variant	ENST00000367429.8
15	rs121913053	503756684	1:196740712	1:196740712	A	G/A	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
16	rs121913062	503756693	1:196743552	1:196743552	T	G/T	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
17	rs460897	501921215	1:196747189	1:196747189	T	C/T	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
18	rs460184	501921185	1:196747207	1:196747207	C	T/C	SNP	dbSNP	pathogenic	missense variant	ENST00000367429.8
19	rs121913059	503756690	1:196747245	1:196747245	T	C/T	SNP	dbSNP	pathogenic~risk factor	missense variant	ENST00000367429.8

Table 4 List of pathogenic SNPs in primary CFH transcript variant (ENST00000367429.8)

Biomart was used to corroborate our prior Ensembl findings. Table 5 below from BioMart yielded 14 rows (as opposed to 18) because it does not include a separate row for each possible allele variant. You can see in the Ensembl Table 4 that rs121913058 has 4 rows and rs121913056 has 2 rows, therefore it is not surprising that the BioMart table (table 5) contains 4 less rows. It should also be noted that there are many categories of clinical significance to investigate besides those chosen here, thus additional queries (adding/subtracting clinical significance categories for example) would likely change the number of SNPs found.

1	Variant name	Variant source	Variant alleles	Clinical significance
2	rs387906550	dbSNP	T/G	pathogenic
3	rs121913058	dbSNP	G/A/C/T	pathogenic
4	rs121913054	dbSNP	G/A/T	pathogenic
5	rs796052138	dbSNP	AGA/-	pathogenic
6	rs1061170	dbSNP	C/T	benign,pathogenic,risk factor
7	rs121913061	dbSNP	C/T	pathogenic
8	rs121913056	dbSNP	T/A/G	pathogenic
9	rs121913052	dbSNP	T/C	pathogenic
10	rs1131690796	dbSNP	A/-	pathogenic
11	rs121913053	dbSNP	G/A	pathogenic
12	rs121913062	dbSNP	G/T	pathogenic
13	rs460897	dbSNP	C/T	pathogenic
14	rs460184	dbSNP	T/C	pathogenic
15	rs121913059	dbSNP	C/T	pathogenic,risk factor

Table 5 Identification of pathogenic SNPs within our region of interest using BioMart

Through the BioMart Human Structural Variants database, additional information was obtained concerning CNVs in the *CFH* gene region. Using chromosome 1 and the *CFH* gene coordinates as filters, 520 entries were found. When the "pathogenic" clinical significance is added, however, 12 CNVs remain (Table 6).

1	Structural variant name	Chromosome/scaffold position start (bp)	Chromosome/scaffold position end (bp)	Structural variant type
2	nsv2772868	914087	248930485	copy_number_variation
3	nsv2768399	120836007	248938897	copy_number_variation
4	nsv2770146	120836007	248938897	copy_number_variation
5	nsv532616	179032905	199724897	copy_number_variation
6	nsv2770915	179042179	199053630	copy_number_variation
7	nsv2776107	179104251	200223137	copy_number_variation
8	nsv2778735	179444344	201795609	copy_number_variation
9	nsv530570	187143981	224299417	copy_number_variation
10	nsv530429	189034483	199615866	copy_number_variation
11	nsv529096	195514309	197896494	copy_number_variation
12	nsv2770490	195514309	248918801	copy_number_variation
13	nsv1197541	196743746	196828416	copy_number_variation

Table 6 Identification of CNVs in the *CFH* gene using BioMart

Galaxy

Galaxy was used to compare the SNP and CNV data found in Ensembl to that of the UCSC gene browser. Interestingly, the information obtained through the UCSC Table Browser in Galaxy was not identical to that which was found in Ensembl. While Ensembl reported 26,340 variants in the ENST00000367429.8 transcript, Galaxy reported 23,673. However, there are many reasons why these numbers may differ, one of which may be related to what external sources/databases UCSC and Ensembl use to gather their variant information or what version the information was pulled from. When using the same search criteria as what was previously described as well as using the Variation group and DGV Struct Var track, Galaxy revealed 53 CNVs. This difference however may be due to the fact that they are not limited on pathogenicity and it is not easy to limit the results further using this tool.

IGV

IGV was used to visualize our genomic area of interest with known SNPs (Figure 12). The boxed region is the SNP of interest, rs1061170. The closest SNP upstream is 3 base pairs away and has the rsID 757876596 while the closest SNP downstream is 11 base pairs away and has the rsID 199705026. All three SNPs are missense variants. However, you can see as a comparison in the same image that if you look at just the Common SNPs(1.4.2) track, there are only two SNPs found in exon 9 of chromosome 1, one of which is rs1061170.

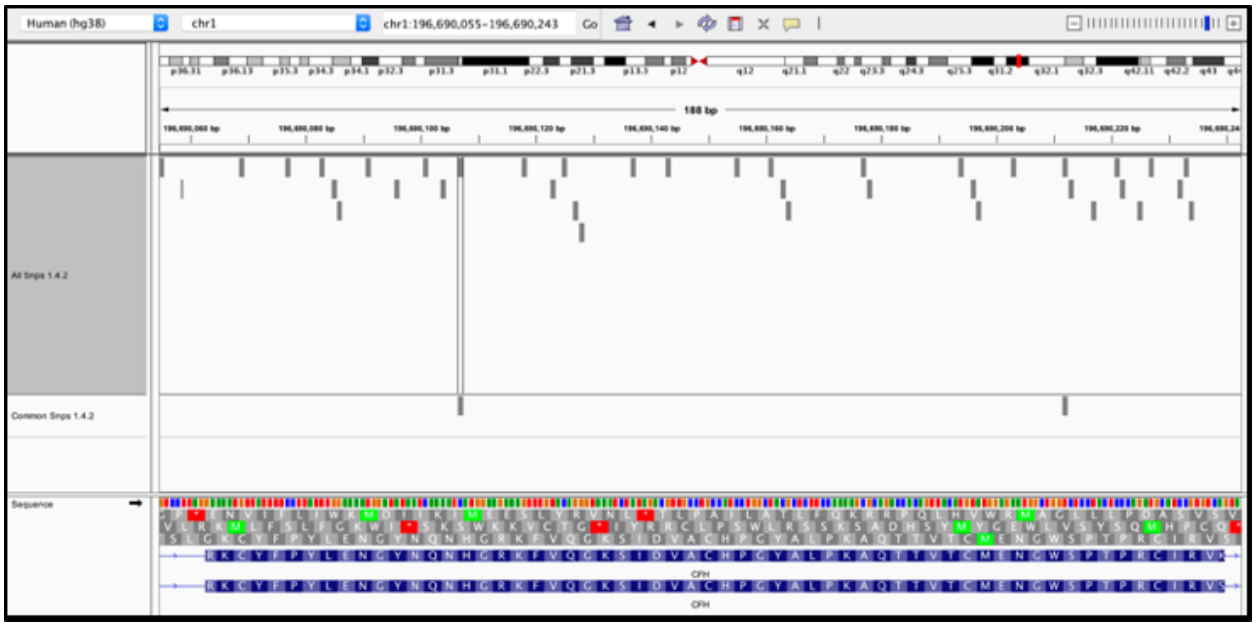


Figure 12 Integrative Genomics Viewer (IGV) visualization of *CFH* gene & known SNPs

Although the Genomic Structural Variation (DGV) track was not available for the hg38 genome in IGV, it was possible to download the 53 CNVs/variants that were found through the Galaxy search performed previously. Although you can not see all of them, you can see in Figure 13 that there are many more found on the 3' end of the gene as opposed to the 5' end. The variants tend to be longer toward 3' end as well. However, it should be noted that the database only included structural variants found in healthy controls.[17]

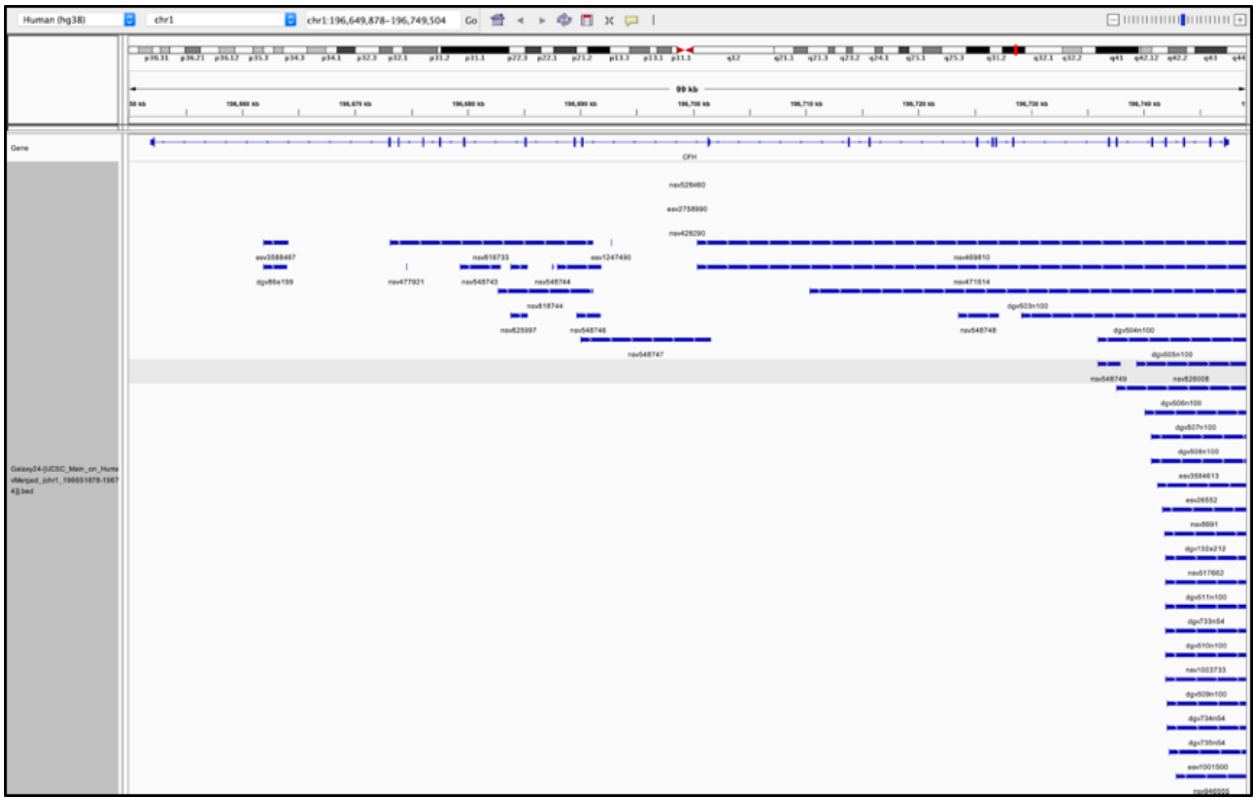


Figure 13 Integrative Genomics Viewer (IGV) visualization of CNVs found within the *CFH* gene.

Although this was useful, we found it easier and more helpful to look at the DGV track in the UCSC Genome Browser. Here you can select the current version of the human genome (hg38) as well as the most recent build of the common SNPs track (151). From Figure 14, you can see similar results as IGV with additional information. Gains in CNVs and/or indels are shown in blue, this mean that more copies were found in the sample than that of the reference genome. Therefore, losses in CNVs and/or indels are shown in red, meaning fewer copies were found in this patient as compared to the reference genome. If the CNV is brown in color, this indicates there are reports of both gains and losses relative to the reference.[18] The second image in figure 14 shows the same information zoomed in on exon 9, where you can see the SNP of

interest highlighted in red. You can more clearly see that 8 CNV/indels overlap this region: 5 show gains (blue), 1 shows loss (red) and 2 show both loss and gain (brown).

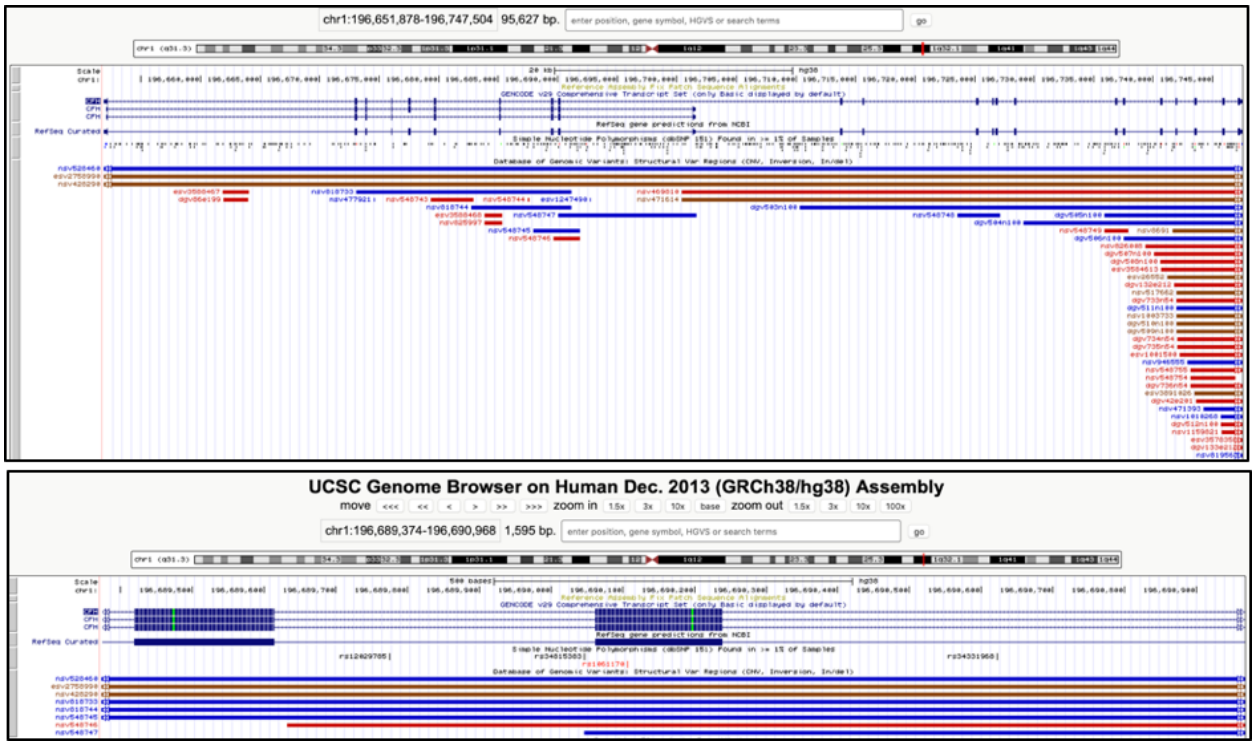


Figure 14 CNVs within the CFH gene as identified by the UCSC Genome Browser

Galaxy, BioMart, IGV, Ensembl and UCSC Genome Browser are all extremely useful tools and are complementary in many ways. IGV seems a little limited as the tracks for the current version of the human annotated genome (hg38) are sparse and they can be more easily accessed in the UCSC Genome Browser.

Promoter and 5'UTR

From the Eukaryotic Promoter Database, we found the predicted promoter region, TSS, start codon, TATA-box, GC-box, and CCAAT-box, all of which can be visualized in Figure 15. No promoter-related polymorphisms have been identified in association with AMD, the primary pathology associated with rs1061170. However, a few CFH promoter polymorphisms have been identified as linked to AHUS (atypical hemolytic uremic syndrome) and membranoproliferative glomerulonephritis type II.[19]

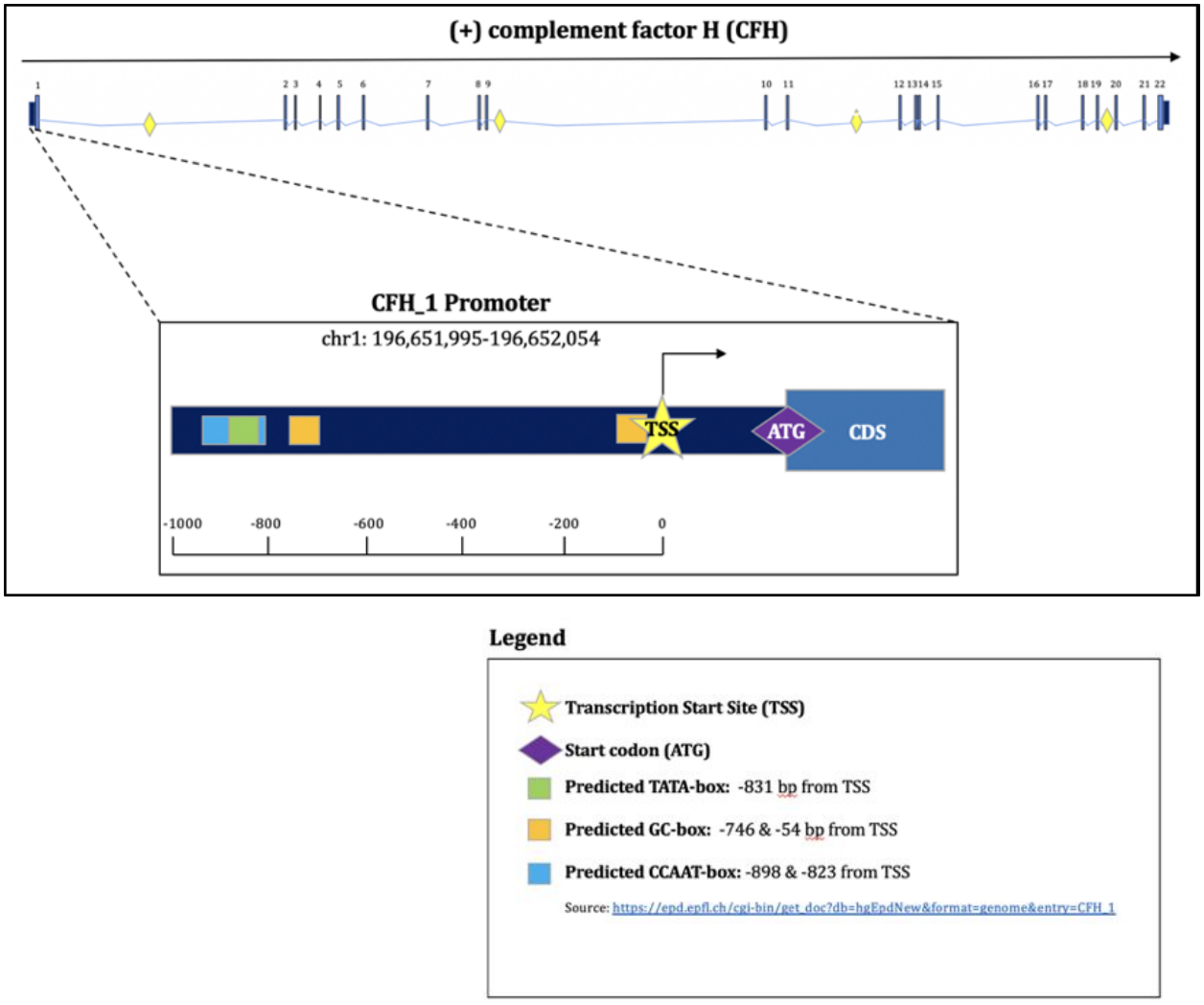


Figure 15 Pictorial representation of the CFH gene highlighting the promoter region and regulatory elements.

3'UTR and miRNA Factors

miRNAs are short non-coding RNAs that act on gene expression post-transcriptionally by (generally) binding to the 3' UTR of mRNA and destabilize or silence the transcript, thus reducing protein translation. miRNA specifically binds to complementary RNA sequences in the 3' UTR of mRNA to account for specificity. miRNAs have been shown to have significant effect on the regulation of CFH in both age-related macular degeneration as well as Alzheimer's disease. In the human brain and retina, CFH is transcribed from a single copy of the CFH gene and creates a large transcript about 4100 nucleotides long. There are three main miRNAs that affect CFH (that we know of): miRNA-125b, miRNA-146a, and miRNA-155. All of these miRNAs recognize an miRNA regulatory control (MiRC) region in the 3' UTR of the CFH gene (Figure 16). By binding to the 3' UTR region, they inhibit the translation of the CFH mRNAs.

These miRNAs contribute to CFH deficiency and increases inflammatory neurodegeneration by down regulating CFH expression leading to deficiencies within the immune system.[20] Deficits in CFH lead to overactivation of the complement pathway in perfectly healthy cells. Overactivation of the complement pathway can often lead to autoimmunity, tissue damage, and/or chronic inflammation. Low levels of CFH expression have been linked to degeneration in both brain and retinal tissues.[20] CFH is a very important regulator in the RCA pathway (regulator of complement activation) and when expressed in low levels, leads to an immune response as well as proinflammatory signals.[20]

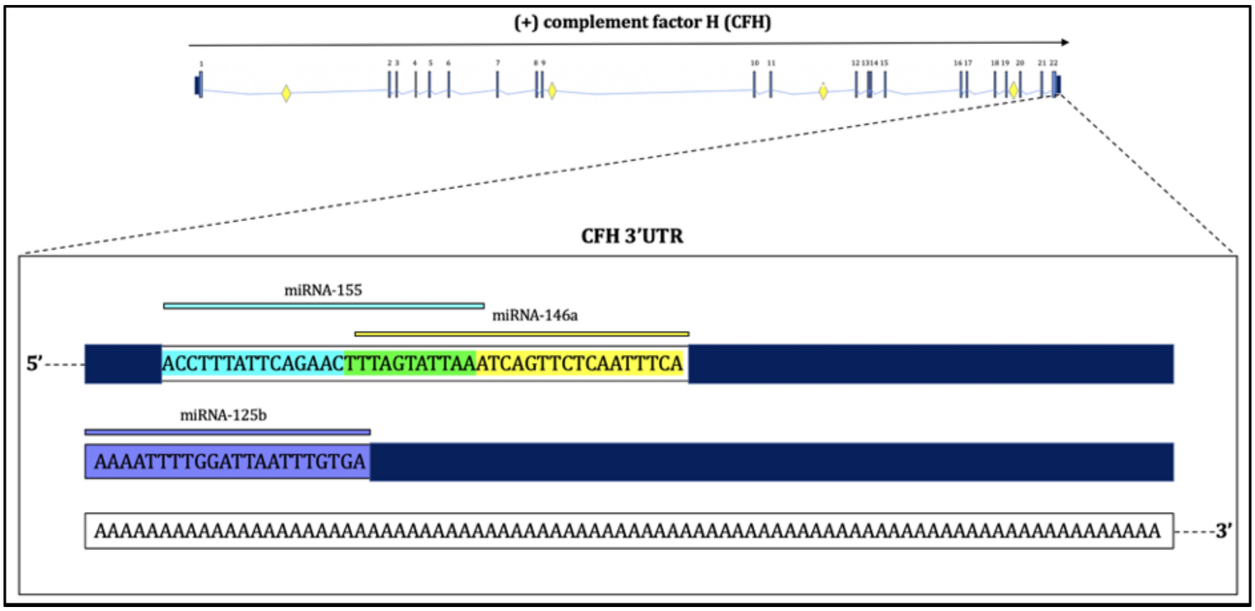


Figure 16: Positions of miRNAs within the 3'UTR of the CFH gene locus.

ENCODE Histone Data Relevant to CFH

We chose to take a closer look at histone modification H3K27Ac in regards to the CFH gene as it is a key histone marker for transcriptional activation, or at least an activation of enhancer properties. This, in conjunction with adequate data of this particular histone modification in the cell lines found in the ENCODE database made this histone modification appropriate to visualize. High levels of H3K27Ac are found at the TSS of the CFH gene at exon 1. There are also other lower levels of H3K27Ac throughout the rest of the CFH gene. The H3K27Ac marks present throughout the CFH gene indicate transcriptional activation, or at least access to the chromosomal region (Figure 17). H3K27Ac is also known to increase transcription of genes by blocking any repressive histone marks such as H3K27me3. High levels of H3K27Ac at the TSS/exon1/enhancer-rich region of CFH also indicates that the enhancers are active as opposed to just being primed, which contain H3K4me1.[21] The presence of H3K27Ac and active enhancers may also indicate high levels of transcription for the entire gene as enhancers can oftentimes act upon regions far from their actual location.

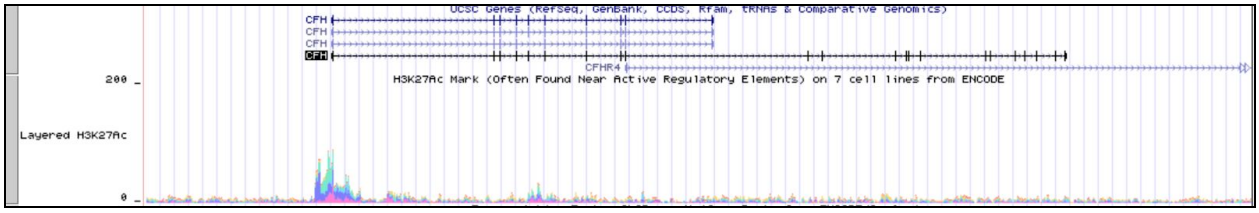


Figure 17 UCSC Genome Browser visualization of the H3K27Ac histone modification in a number of cell lines from the ENCODE database.

H3K27Ac signals are found in the following cell lines: H1-hESC, HUVEC, and HSSM in exon 9 where SNP rs1061170 is located. H3K27Ac, as mentioned previously is often an indicator of transcriptional activation, enhancer activation, or even a blocker for repressive markers. These H3K27Ac patterns show that there is a high likelihood of elevated expression of CFH in many of the "normal" cell lines included in the ENCODE database (Figure 18). From this, we can visualize how a normal CFH gene would be transcriptionally regulated via histone modifications. However, without ChIP-seq data from a patient with our particular SNP in a relevant tissue, we can't know how this modification changes CFH expression. We can, however, hypothesize that in addition to its other effects: our SNP may also cause histone deacetylation of H3K27 around the CFH gene as a whole, which may lead to the dysregulation of the complement system in the ocular environment.

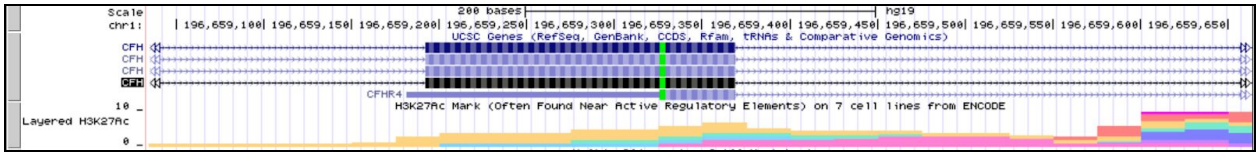


Figure 18 UCSC Genome Browser visualization of the H3K27Ac modification zoomed in to the region of exon 9 and our SNP of interest (rs1061170).

Next Generation Sequencing Process and Workflow

Upon uploading the control and AMD samples into Galaxy using the associated accession numbers, FASTQC was run to obtain information on read quality. As you can see in the Figures 19 and 20, the control sample contains reads ranging from 35 to 76 base pairs in length whereas the AMD sample contains reads that are 100 base pairs in length.

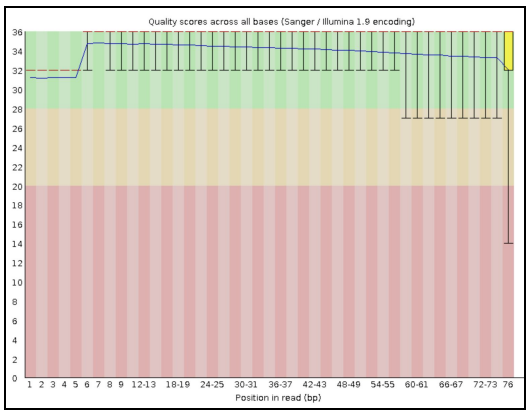


Figure 19. FASTQC on Control sample

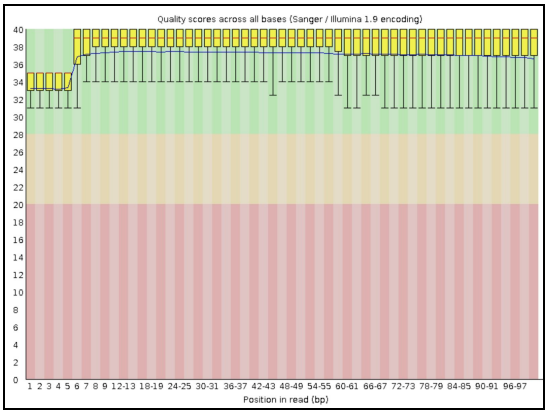
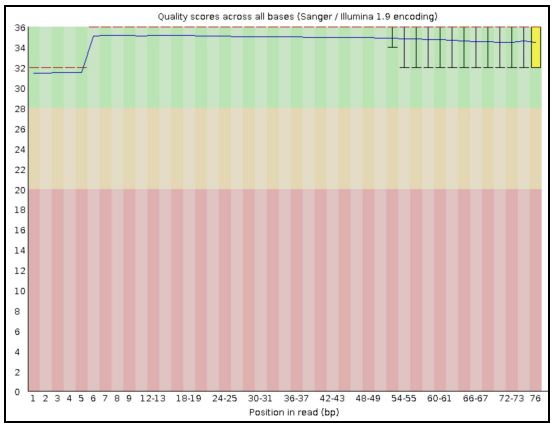


Figure 20. FASTQC on AMD sample 1

Additionally, the control sample contained 124,772,162 total sequences and the AMD sample contained 125,288,726 total sequences. Because the forward and reverse reads were interlaced together, we used the FASTQ-deinterlacer tool on the paired end reads. This separated the paired reads into individual files so we could run the Trimmomatic tool. Although the overall quality of the reads looked pretty good, we ran the Trimmomatic tool and reran the FASTQC tool to ensure high quality reads for both the forward and reverse reads from the control and AMD samples (Figures 21-24).



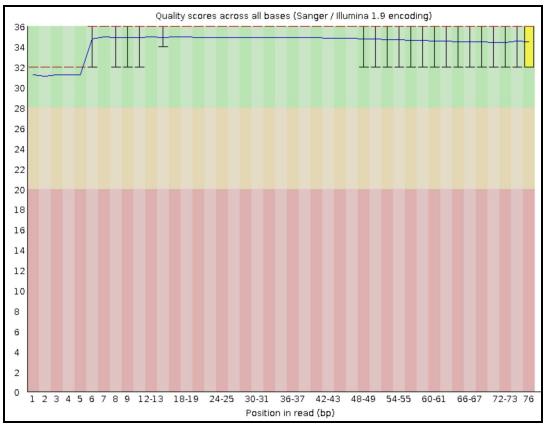
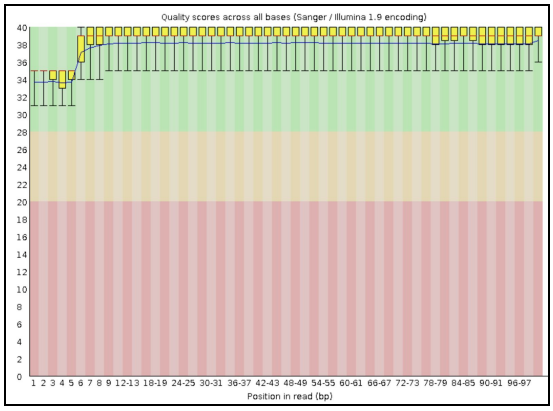


Fig 21. FASTQC on trimmed Control forward reads **Fig 22.** FASTQC on trimmed Control reverse reads



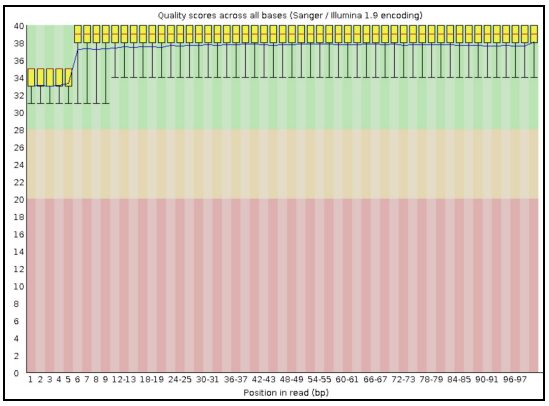


Fig 23. FASTQC on trimmed AMD 1 **Fig 24.** FASTQC on trimmed AMD 1 reverse reads forward reads

In order to understand more about the reads, HISAT was used to find the positional read locations within the human genome (hg38). Because our samples are a result of RNA-seq, we chose not to use Bowtie2 or BWA as these tools aren't able to accurately map reads that span splice junctions. The following parameters were used when running HISAT on both the control and AMD samples: reference genome - hg38, paired end reads - file 1 indicates the forward reads, file 2 indicates the reverse reads. The hg38 genome was used for alignment as it is the most up to date build of the human genome (2013) and our samples were collected in 2017/2018. Following alignment, the FreeBayes tool was used to detect genetic variants within the two samples. We used the HISAT2 BAM files as input, the hg38 genome as the reference and limited the region to chr1 (region: 0 - 248,956,422) for more manageable results. The control VCF file contained 76,225 lines/variants and the AMD VCF output file contained 67,537 lines/variants.

To get a better idea of what variants were found in the CFH gene and more specifically to see if our SNP of interest, rs1061170, was found in the sample from the AMD patient, we loaded the VCF files into IGV. Interestingly, you can see variants unique to the control and AMD samples within the CFH gene, however, neither sample contained the rs1061170 SNP in exon 9 at base pair 196,690,107 (Figure 25).

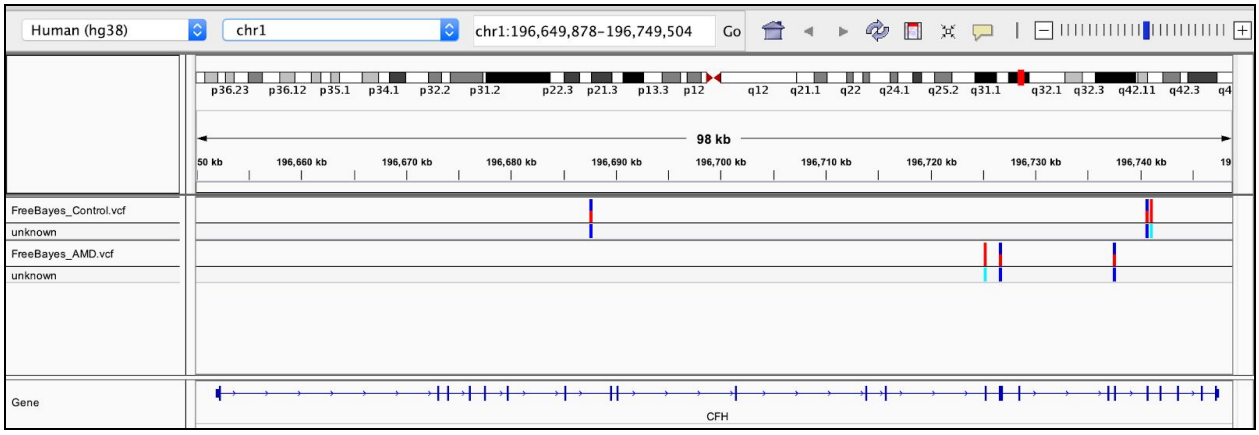


Figure 25 Visualization of SNPs found in the NGS control sample and AMD 1 sample within the CFH gene.

In an attempt to find an AMD sample with our SNP of interest, we ran another sample. The second AMD sample was from the same individual but came from the left eye. It was described as "late AMD" versus "early AMD". After following the same protocol/pipeline steps previously described, the variants found as a result of the FreeBayes tool were uploaded to IGV. Unfortunately however, as can be seen in Figure 26, there were no SNPs found in exon 9 in the second AMD sample either.

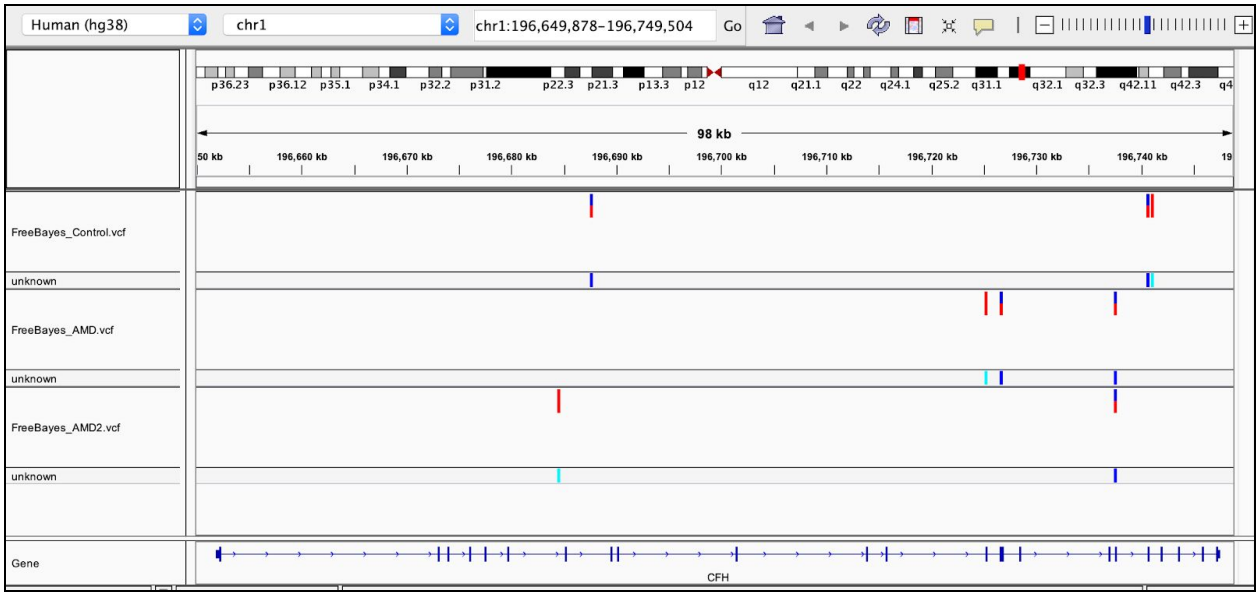


Figure 26 Visualization of SNPs found in the NGS control sample and both AMD samples within the CFH gene.

Interestingly though, both AMD samples contained the same SNP located on chromosome 17 at base pair 196,737,512 and according to ClinVar, SNP rs55752475 is associated with macular degeneration.[22]

DISCUSSION

While it is disappointing that our analyzed samples did not contain our SNP of interest, it is not necessarily surprising. Other studies have found that our SNP was found to increase the likelihood of a patient to have AMD, but have also made it clear that this SNP is not necessarily the sole causative agent of AMD. Although SNP rs1061170 is not present in our selected samples, we do corroborate the idea that other SNPs are present in the AMD patient versus the control patient and that they may play a role in the development of AMD.

In our investigation of population genetics, allele frequencies have been found to correlate with genetic diversity amongst individuals and the rare population-specific variants have been shown to indicate strong functional effects.[23] Age-related macular degeneration (AMD) is an example. AMD is the leading cause of irreversible blindness in both Western and Asian industrial countries and is found to be less common in African-American and Hispanic/Latino populations.[24,25] As previously discussed, SNP rs1061170 was found to be associated with AMD disease susceptibility. While rs1061170(T) allele encodes the more common Tyr (Y), generally rarer allele rs1061170(C) encodes the His (H).[26] However, further analyses revealed that the relationship between disease susceptibility and AMD risk differs by geographical population.

Zareparsi et al. for example found that after studying 616 Caucasian individuals with AMD (and 275 controls), the C allele was found at a significantly higher frequency in patients with AMD than without AMD. More specifically, individuals carrying at least one copy of the C have a 4.36-fold increased risk of AMD. Homozygous individuals (CC) exhibit a 5.52-fold increase risk of AMD and this genotype was more commonly found in patients with a family history of AMD.[27] In a follow up study, 84 SNPs found in and around CFH were analyzed in 726 patients with AMD (268 controls). Interestingly, 20 SNPs showed a stronger association with AMD susceptibility than rs1061170. The three SNPs that resulted in the highest correlation with AMD can be seen in the Figure 6 (rs2274700, rs1410996, rs7535263). Therefore, their results indicate that rs1061170 is not the only major determinant of AMD risk. They go on to propose that this SNP may be in linkage disequilibrium with nearby alleles.[28]

While the rs1061170 or Y402H variant has been shown to be associated with AMD susceptibility in Caucasians, in some studies, the strong association was not found in Asian populations.[25] Gotoh et al. performed a comparable study to Zareparsi et al. in Japanese patients, however, the reported risk of carrying at least one C allele in Japanese patients was as low as 0.04 ($\chi^2 = 3.19$, $P_{corr} = 0.423$) as compared to $\chi^2 = 110.96$, $P < 1 \times 10^{-24}$ in Caucasians. The population of Japanese patients in which these numbers were obtained were strictly selected cases of AMD in late stages which is when rs1061170 has shown the most significant association. In addition, no other

AMD associated SNPs within CFH in Caucasians were shown to associate with AMD in this group of Japanese patients.[25]

Interestingly, a meta-analysis using 76 published studies (27418 AMD patients and 32843 controls) was completed by Maugeri et al. to determine the strength of the association between rs1061170 and AMD.[29] The analysis was stratified by stage of disease and ethnicity. As expected, the results indicated a significant association between rs1061170 and AMD and the strength of association is reduced as you move geographically from West to East. In European populations there was a strong association between rs1061170 and AMD and the risk was 2.5-fold in individuals with at least one copy of the risk allele.[29] Interestingly, while many individual studies (like the one above) have found a lack of association between rs1061170 and AMD risk in Asian populations, others have found that rs1061170 does correlate with disease susceptibility in Asians.[30,29]

When analysis by both AMD subtype and ethnicity was performed, the following results were obtained: In Caucasians, the risk was lower for early AMD compared to advanced AMD, while in Asians, rs1061170 was significantly associated with advanced AMD but not early AMD. More specifically, in caucasians, the presence of the C allele resulted in 2.9 fold increased risk of dry AMD and 2.5-fold increased risk of wet AMD. In Asians, however, the presence of the C allele resulted in a 2.2-fold increased risk of wet AMD but it wasn't found to be associated with dry AMD.[29]

CONCLUSION

While verifying the presence of SNP rs1061170 in our NGS samples would help validate our hypothesis, our computing resources were limited. Access to more high performance computing (HPC) resources would enable a more comprehensive review of samples and corroboration of SNP rs1061170 involvement in AMD cases (as already documented in numerous studies). As it stands, however, we were only able to run the three samples shown that did not display the presence of our SNP. Therefore, additional NGS analysis is of great interest in the future. Not only would it be beneficial to study the frequency of SNP rs1061170 in patients displaying AMD compared to those that do not display AMD, but it would also be interesting to compare the frequency of SNP rs1061170 amongst individuals with early versus late onset AMD or individuals with wet versus dry AMD.

References

1 Yang HJ *et al*. (2015). "Vision from next generation sequencing: Multi-dimensional genome-wide analysis for producing gene regulatory networks underlying retinal development, aging and disease." *Prog Retin Eye Res.* 46: 1–30. doi: 10.1016/j.preteyeres.2015.01.005.

2 Klein RJ *et al*. (2005). "Complement factor H polymorphism in age-related macular degeneration." *Science*. 308(5720): 385-389. doi: 10.1126/science.1109557.

3 "CFH Gene." *Genetic Home Reference*. https://ghr.nlm.nih.gov/gene/CFH#resources.

4 Rodríguez S *et al*. "The human complement factor H: functional roles, genetic variations and disease associations". *Mol Immunol*. 41(4): 355-367. ISSN 0161-5890.

5 "COMPLEMENT FACTOR H; CFH." *OMIM - Online Mendelian Inheritance in Man*, www.omim.org/entry/134370.

6 Narendra U *et al*. (2009). "Genetic analysis of complement factor H related 5, CFHR5, in patients with age-related macular degeneration." *Mol Vis.* 15: 731-736. PMID: 19365580.

7 Cantsilieris S et al. (2018). "Recurrent structural variation, clustered sites of selection, and disease risk for the complement factor H (CFH) gene family." *PNAS USA*. 115(19): 4433-4442. doi: 10.1073/pnas.1717600115.

8 Toomey *et al*. (2018). "Complement factor H in AMD: Bridging genetic associations and pathobiology." *Prog Retin Eye Res*. 62: 38-57. doi: 10.1016/j.preteyeres.2017.09.001.

9 Harrison R and Morikis D. (2019). "Molecular Mechanisms of Macular Degeneration Associated with the Complement Factor H Y402H Mutation." *Biophys J*. 116(2): 215-226. doi: 10.1016/j.bpj.2018.12.007.

10 Mohamad *et al*. (2018). "Analysis of the association between CFH Y402H polymorphism and response to intravitreal ranibizumab in patients with neovascular age-related macular degeneration (nAMD)." *Bosn J Basic Med Sci*. 18(3): 260-267. doi: 10.17305/bjbms.2018.2493.

11 https://www.ncbi.nlm.nih.gov/snp search = CFH[All Fields]

12 Swaroop et al. (2009). “Unraveling a Multifactorial Late-Onset Disease- From Genetic Susceptibility to Disease Mechanisms for AMD.” Annu Rev Genomics Hum Genet. 10: 19-43. doi: 10.1146/annurev.genom.9.081307.164350

13 Toomey CB *et al*. “Regulation of Age-Related Macular Degeneration-like Pathology by Complement Factor H.” *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, 19 May 2015, doi.org/10.1073/pnas.1424391112.

14 Clark SJ. (2014). “Identification of Factor H–like Protein 1 as the Predominant Complement Regulator in Bruch’s Membrane: Implications for AMD.” *J Immunol*. 193(10): 4962-70. doi: 10.4049/jimmunol.1401613.

15 Harrison RES *et al*. (2019). “Molecular Mechanisms of Macular Degeneration Associated with the Complement Factor H Y402H Mutation.” *Biophys J.* 116(2): 215-226. doi: 10.1016/j.bpj.2018.

16 “Gene: CFH ENSG00000000971.” Ensembl, Ensembl, uswest.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Table?db=core;g=ENSG00000000971;r=1:196651878-196747504.

17 “Genomic Variants in Human Genome (Build GRCh38: Dec. 2013, hg38): 800 Kbp from chr7:71,890,181..72,690,180.” *Database of Genomic Variants*, dgv.tcag.ca/gb2/gbrowse/dgv2_hg38/.

18 “ Database of Genomic Variants: Structural Var Regions (CNV, Inversion, In/Del) (nsv528460).” *UCSC Genome Browser*, genome.ucsc.edu/cgi-bin/hgc?hgsid=715506389_Xtxd1WnGWrF1deYlwBqLqPFcyAlw&c=chr1&l=196689639&r=196690702&o=196573436&t=197477572&g=dgvMerged&i=nsv528460.

19 Fraczek LA *et al*. (2011). “c-Jun and c-Fos regulate the complement factor H promoter in murine astrocytes.” *Mol Immunol.* 49(1-2): 201-10. doi: 10.1016/j.molimm.2011.08.013.

20 Lukiw WJ *et al*. “Common micro RNAs (miRNAs) target complement factor H (CFH) regulation in Alzheimer's disease (AD) and in age-related macular degeneration (AMD).” *International journal of biochemistry and molecular biology* vol. 3,1 (2012): 105-16.

21 Creyghton MP *et al*. "Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State." *Proceedings of the National Academy of Sciences*, vol. 107, no. 50, 2010, pp. 21931–21936., doi:10.1073/pnas.1016071107.

22 "NM_000186.3(CFH):C.2637A>G (P.Gly879=) Simple - Variation Report - ClinVar - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/clinvar/variation/294505/.

23 Geerlings MJ *et al*. "Geographic distribution of rare variants associated with age-related macular degeneration."*Molecular vision*. vol. 24 75-82. 27 Jan. 2018.

24 Yu J *et al*. "Biochemical Analysis of a Common Human Polymorphism Associated with Age-Related Macular Degeneration†." *Biochemistry*, vol. 46, no. 28, 2007, pp. 8451–8461. doi:10.1021/bi700459a.

25 Gotoh N *et al*. "No Association between Complement Factor H Gene Polymorphism and Exudative Age-Related Macular Degeneration in Japanese." *Human Genetics*, vol. 120, no. 1, 2006, pp. 139–143. doi:10.1007/s00439-006-0187-0.

26 "rs1061170." *SNPedia*, www.snpedia.com/index.php/Rs1061170

27 Zareparsi S *et al*. "Strong association of the Y402H variant in complement factor H at 1q32 with susceptibility to age-related macular degeneration." *American journal of human genetics*vol. 77,1 (2005): 149-53. doi:10.1086/431426

28 Li M *et al*. "CFH Haplotypes without the Y402H Coding Variant Show Strong Association with Susceptibility to Age-Related Macular Degeneration." *Nature Genetics*, U.S. National Library of Medicine, Sept. 2006, www.ncbi.nlm.nih.gov/pmc/articles/PMC1941700/.

29 Maugeri A *et al*. "The Association between Complement Factor H rs1061170 Polymorphism and Age-Related Macular Degeneration: a Comprehensive Meta-Analysis Stratified by Stage of Disease and Ethnicity." *Acta Ophthalmologica*, vol. 97, no. 1, 2018, doi:10.1111/aos.13849.

30 Kondo N, Bessho H, Honda S, Negi A (2011): Complement factor H Y402H variant and risk of age‐related macular degeneration in Asians: a systematic review and meta‐analysis. *Ophthalmology* **118**: 339– 344.